Geoderma Regional 10 (2017) 154–162

Contents lists available at ScienceDirect

Geoderma Regional

journal homepage: www.elsevier.com/locate/geodrs

Spatial prediction of major soil properties using Random Forest techniques - A case study in semi-arid tropics of South India

S. Dharumarajan[a,*], Rajendra Hegde[a], S.K. Singh[b]

[a] ICAR-National Bureau of Soil Survey and Land Use Planning, Regional Centre, Hebbal, Bangalore 560024, India
[b] ICAR-National Bureau of Soil Survey and Land Use Planning, Amaravati Road, Nagpur-440033, India

ARTICLE INFO

Keywords:
Soil properties
Digital soil mapping
Random Forest model
Prediction
Validation

ABSTRACT

The purpose of the study is to map the spatial variation of major soil properties in Bukkarayasamudrum mandal of Anantapur district, India using Random Forest model. The study area is divided into different Physiographic Land Units (PLU) based on landform, landuse and slope. Random Forest model (RFM) was developed based on field survey data of 116 surface samples (0–30 cm) representing all major PLU units of the study area. RFM is neither sensitive to over fitting nor to noise features and has capacity to handle large datasets. High resolution satellite imagery (IRS LISS IV data- 3 bands), terrain attributes such as elevation, slope, aspect, topographic wetness index, topographic position index, plan & profile curvature, Multi-resolution index of valley bottom flatness and Multi-resolution ridge top flatness, Vegetation factors like NDVI, EVI and land use land cover (LULC) are used as covariates along with legacy soil data of 1:50,000 scale. The predicted organic carbon, pH and EC ranged from 0.24–1.03%, 6.9–9.0, 0.11–0.97 dsm^{-1} respectively. The model performance was evaluated based on Coefficient of determination (R^2) and Lin's Concordance coefficient (CCC). The model performed well with R^2 and CCC values of 0.23 and 0.38 for SOC, 0.30 and 0.37 for pH, and 0.62 and 0.70 for EC respectively. Variable importance ranking of RFM model showed that EVI and NDVI are the most important predictors for organic carbon whereas drainage and NDVI for EC and pH respectively. This technique can be applied to similar landscapes with more observations to refine the spatial resolution of soil properties.

1. Introduction

Most of the environment modeling work requires spatially continuous and quantitative soil information especially at larger scale (Gessler et al., 1996; Minasny et al., 2008; Hartemink and McBratney, 2008). Such information is always not available at the required scale (Lagacherie et al., 1995; McBratney et al., 2003; Greve et al., 2012) and mapping at high accuracy is always challenging, time consuming and costly (Vågen et al., 2016).

In recent past, digital soil mapping (DSM) techniques have become more popular among the natural resource mappers as it offers the solution for spatial prediction of soil attributes in quicker way (Lagacherie and McBratney, 2007). DSM mapping is based on soil forming factors models like CLORPT (Jenny, 1941) and *SCORPAN* method which describes the soil formation as a function of climate(c), organisms (o), topography (r), parent materials (p), age (a) and spatial location (McBratney et al., 2003). DSM techniques can be applied for both prediction of quantitative outputs like sand, silt, clay, pH, electrical conductivity and qualitative outputs like soil taxonomic units using regression algorithm (Wiesmeier et al., 2011; Akpa et al., 2014) and classification algorithm (McBratney et al., 2003; Hastie et al., 2009; Kidd et al., 2014) respectively.

In DSM, the interrelationships between the soil and environmental covariates are brought out to predict the soil information by using different statistical and geostatistical models. For instance, Organic carbon is predicted by multiple linear regression (Powers and Schlesinger, 2002; Thompson et al., 2006), neural networks (Mansuy et al., 2014; Malone et al., 2009), tree models (Henderson et al., 2005), generalized linear models (McKenzie and Ryan, 1999) and regression kriging (Hengl et al., 2004; Simbahan et al., 2006).

Use of Machine learning algorithms in digital soil mapping is another approach to modeling soil classes and soil properties (Huang et al., 2002; Rogan et al., 2003).These algorithms are faster and more efficient in classification (Foody, 2002; Friedl and Brodley, 1997). Machine language techniques predict the soil property for unvisited location using interrelationship with the environmental covariates such as digital elevation models (DEMs) (McBratney et al., 2000), climatic parameters (Akpa et al., 2014), remote sensing imageries (Odeh and

* Corresponding author.
E-mail address: Dharumarajan.S@icar.gov.in (S. Dharumarajan).

http://dx.doi.org/10.1016/j.geodrs.2017.07.005
Received 12 April 2017; Received in revised form 22 July 2017; Accepted 24 July 2017
Available online 26 July 2017
2352-0094/

McBratney, 2000), and legacy soil information (Akpa et al., 2014).

Some of the major machine language techniques used in DSM are classification and regression trees (Breiman et al., 1984), k-nearest neighbour (Mansuy et al., 2014), multinomial logistic regression (Kempen et al., 2009), logistic model trees (Giasson et al., 2006), Support vector machine (Kovačevic et al., 2010; Priori et al., 2014) and Random Forest model (Vågen et al., 2016). Out of four different machine language techniques studied by Rodriguez-Galiano and Chica-Rivas (2014), Random Forest model was found to be most accurate and more robust in noise and data reduction. It has capacity to handle both quantitative and qualitative datasets. The potential of RFM in digital soil mapping has been demonstrated by several authors (Grimm et al., 2008; Hastie et al., 2009; Wiesmeier et al., 2011; Vågen et al., 2016; Sreenivas et al., 2016).

Despite the wide spread of various machine learning algorithms used in DSM, limited work is available in Indian subcontinent (Sreenivas et al., 2016). In this context, the present study was carried out is to produce a fine resolution map for major soil properties, Organic carbon, pH and EC in Bukkarayasamudrum mandal of Anantapur district representing semi-arid tropics of south India using Random Forest model techniques.

2. Material and methods

2.1. Description of the study area

Bukkarayasamudrum mandal is located between 13°37′ 51″ and 14°48′ 09″ N latitudes and 77°33′ 47″ and 77°47′45″ E longitudes in Anantapur district of Andhra Pradesh, India (Fig. 1). The total area of the district is 24,808 ha. The climate of Bukkarayasamudrum mandal is warm and classified as hot and arid. The mean minimum and maximum temperatures are 22.9° and 34 °C with average rainfall of 556 mm. Major geology is granite- gneisses. Quartz, feldspar and mica, are the major mineral composition of granite and gneissic rock types. The elevation ranges from 295 to 595 m MSL. The major part of mandal has nearly level to very gently sloping with a slope ranging from 1 to 3%. The major soils of Bukkarayasamudrum mandal are moderately deep (75–100 cm) gravelly red clayey soils (Clayey-skeletal, mixed, isohyperthermic Typic Haplargids) followed by shallow (25–50 cm) gravelly red clayey soils (Clayey-skeletal, mixed isohyperthermic, Lithic Haplargids) and deep (100–150 cm) black clayey soils (Fine, mixed isohyperthermic (cal) Ustic Haplargids). The study area is chronically prone to drought due to its low and unpredictable nature of length of the growing period (LGP). The LGP of Bukkarayasamudrum mandal is 11 weeks starting from third week of August and ending in the last week of October. The net cultivated area is 47.4% of total area. The major rain fed crops are groundnut, pearl millet, sorghum and minor millet while rice, cotton and vegetables are important irrigated crops.

2.2. Soil sampling and analysis

We used physiographic land unit map (PLU) as a base for soil sampling. Physiographic land unit is the assemblage of landform, slope and land use. Landform represents testimony of climatic events whereas slope and land use represent the influence of present climatic conditions on the soil formation. Survey of India toposheets (1:50,000 scale) and Resourcesat-2 LISS IV (Linear imaging self scanner - IV) imagery (5.8 m resolution) were used for preparation of landform and land use map.

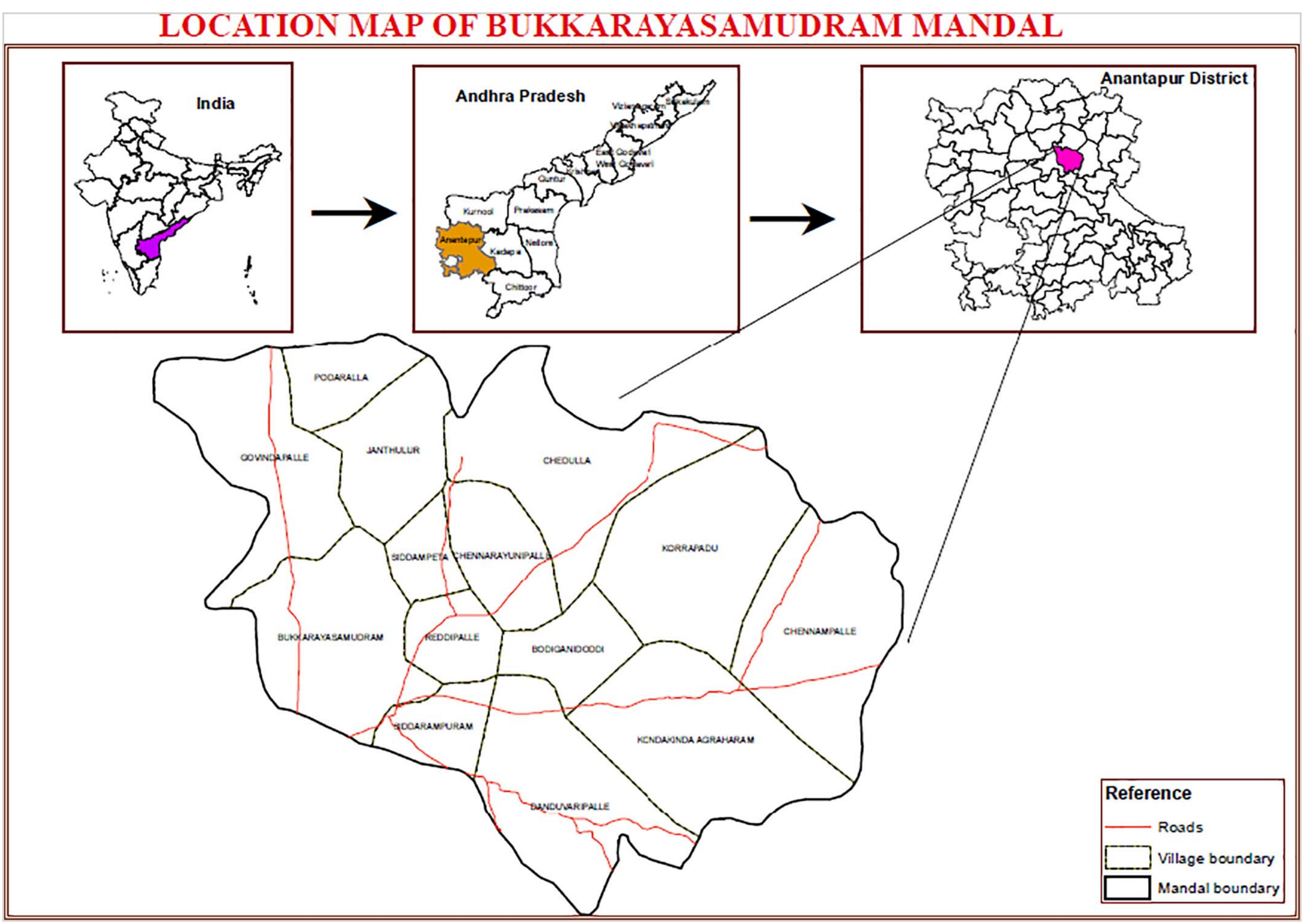

Fig. 1. Location map.